AD-A033 992 NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN D--ETC F/G 5/9
MEASUREMENT OF JOB-PERFORMANCE CAPABILITIES.(U)
DEC 76 E J PICKERING, A V ANDERSON

UNCLASSIFIED NPRDC-TR-77-6 NL

1 OF 1
AD-A
033 992

END
DATE
FILMED
2-15-77
NTIS

MEASUREMENT OF JOB-PERFORMANCE CAPABILITIES

NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA

DECEMBER 1976

NPRDC TR 77-6

December 1976

MEASUREMENT OF JOB-PERFORMANCE CAPABILITIES

Edward J. Pickering
Adolph V. Anderson

Reviewed by
Richard C. Sorenson

Approved by
James J. Regan
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPRDC TR 77-6	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) MEASUREMENT OF JOB-PERFORMANCE CAPABILITIES		5. TYPE OF REPORT & PERIOD COVERED Interim July 1975 - July 1976
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Edward J. Pickering Adolph V. Anderson		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62763N 522-002-03-40
11. CONTROLLING OFFICE NAME AND ADDRESS Navy Personnel Research and Development Center San Diego, California 92152		12. REPORT DATE December 1976
		13. NUMBER OF PAGES 38
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Performance Measurement
Personnel Quality Control
Job-Performance Capabilities

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This Center is planning an Advanced Development effort aimed at the development of a comprehensive system for obtaining and reporting Navy job-performance capability information. In preparation for that effort, a review and an analysis were undertaken of performance measurement techniques that might support the proposed system. This report presents the results of that review and analysis and suggests a general approach to meeting the Navy's requirements for incumbent capability information relative to critical tasks. The approach which

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

is suggested is based upon the use of quality control techniques analogous to those utilized in the manufacturing of industrial products.

FOREWORD

This research and development was conducted in support of Exploratory Development Task Area ZF55.522.002 (Methodology for Development and Evaluation of Navy Training Programs) under the sponsorship of the Chief of Naval Education and Training. It was carried out under Work Unit ZF55.522.002.03.40: Techniques for Measurement of Job Proficiency. The project was initiated by the Navy Personnel Research and Development Center in order to identify, devise, and develop techniques and instruments for the measurement of the job-performance capabilities of the Navy's officers and enlisted men in their individual and team assignments.

J. J. CLARKIN
Commanding Officer

SUMMARY

Problem

In order to ensure that the man part of the Navy's man-machine system is performing as it should, the Navy's personnel managers must have information as to how capable operational personnel are of performing the critical aspects of their assigned jobs. While a number of existing programs provide some performance data, no comprehensive system is currently available which provides personnel managers with all of the performance information they require.

Objective

The Navy Personnel Research and Development Center is developing a comprehensive system for obtaining job-proficiency information and reporting that information to key decision makers. In preparation for that development, the objectives of this effort were: (1) to review and analyze performance measurement techniques that might logically support the proposed system and (2) to provide a suggested general approach to be followed in the development of a job-proficiency assessment system.

Approach

A review and analysis was carried out of the performance measurement research literature. On the basis of that review, a general approach was developed for meeting the Navy's need for information concerning how well individuals can perform critical aspects of their jobs.

Conclusions

Some form of job performance testing appears to be the best source of incumbent capability information relative to critical Navy tasks. However, job performance testing is difficult, expensive, and demands substantial expertise. It is evident that the costs of measuring all aspects of job performance for all Navy incumbents would be prohibitive. It is also evident that where job performance tests have been developed and used, typically insufficient attention has been given to evaluating the quality of the performance measurement instruments themselves.

Recommendations

It is recommended that:

1. The Navy develop and evaluate a prototype proficiency assessment system based on the quality control approach discussed in this report.

2. A series of experimental studies be conducted to provide additional information and guidance on job measurement techniques relative to such questions as objectivity, reliability, validity, and guidelines for use of testing procedures in measurement programs.

CONTENTS

LIST OF TABLES

FIGURE

INTRODUCTION

Problem

The Navy is a very large complex man-machine system. The combined impacts of inflation, changing social values, and a continuing requirement to remain constantly ready to carry out the various aspects of its mission demand that all parts of this system perform effectively and efficiently. To ensure that the man part of the Navy's man-machine system performs as it should, the Navy's personnel managers must have information as to how well individuals can perform the critical aspects of their assigned roles. While a number of existing procedures yield some performance information, many studies conducted over a wide time span have shown that deficiencies exist in key aspects of personnel performance and that supervisors and managers are often unaware of these deficiencies (Anderson, 1963; Anderson & Pickering, 1959a, 1959b; Winchell, Panell, & Pickering, 1976).

Objective

In order to provide all required personnel performance data, a comprehensive system for obtaining and reporting Navy job proficiency information has been proposed (Performance Proficiency Assessment System, Subproject 31 of Z0108 PN). The objectives of this effort were (1) to prepare for that development effort by conducting a review and analysis of performance measurement techniques that might logically support the proposed measurement system and (2) to provide a suggested general approach to meeting the Navy's job performance measurement requirements.

Background

In the past, various techniques (e.g., rating scales, interviews, observational techniques, critical incident techniques, job diaries) have been used to obtain information concerning how well individuals can carry out the critical aspects of their jobs. However, for purposes of providing feedback to decision makers upon which specific corrective action can be taken, none of these techniques can provide the type of precise information which is provided by actually testing individuals on how well they can perform specific, critical tasks. Numerous studies carried out by this Center and other organizations (A. Abrams, 1962; A. Abrams & Klipple, 1965; Ainsworth & Bishop, 1971; Anderson, 1963; Anderson & Pickering, 1959a, 1959b; Branks, 1966; Brock, Wells, & M. Abrams, 1974; Engel, 1970; Klipple & A. Abrams, 1966; Megling & M. Abrams, 1973; Pickering, 1959; Shriver & Foley, 1974; Whipple, Baldwin, Mager, & Vineberg, 1969; Wilson & Mackie, 1952; Winchell et al., 1976; also numerous classified NPRDC reports) have demonstrated the value of performance tests for identifying specific performance strengths and weaknesses. Consequently, such performance tests will form the backbone of the proposed Performance Proficiency Assessment System. Evaluation techniques other than performance testing are generally well-developed and understood and are supported by a voluminous literature. In this effort, no attempt was made to survey that literature. Instead, concentration was focused on the less developed and more pertinent area of performance testing on individual job skills.

REVIEW AND ANALYSIS OF PERFORMANCE MEASUREMENT TECHNQIUES

A job performance test can be defined as one in which one or more individuals are required to accomplish a job-related task under controlled conditions. Such tests offer one of the most direct means of determining whether or not individuals are capable of performing critical portions of their jobs. Capable is underlined to emphasize that such tests measure only capability. Many other factors (e.g., motivation, environment, command structure) may enter into how an individual actually performs when he is on the job. However, if performance capability on a specific critical task is absent, it is certain that job performance on that task will be unsatisfactory.

Properly designed job performance tests provide precise measures of job-skill capabilities. Additionally, they can provide valuable diagnostic information concerning the specific nature of detected deficiencies. However, as compared to paper-and-pencil tests, they are costly to develop and to administer. Consequently, major assessment programs, either within or outside the military, that rely on performance tests as their primary data source are almost nonexistent.

Degree of Simulation

An important consideration in the development of any job performance test is the degree of simulation that it demands. It may be necessary to test on the actual operational equipment; an equipment mock-up may be required; some form of computerized simulation may be appropriate; a pictorial representation of the equipment may be sufficient; or it is possible that a series of written questions may be sufficient and a performance test is not required.

Only a small amount of significant research has been carried out concerning the problem of determining the appropriate simulation requirements for measuring proficiency on individual job skills. Foley (1974) presents an excellent review of research on the effectiveness of paper-and-pencil substitutes for job performance tests in the electronic maintenance area. Table 1 summarizes the studies reviewed that compare the relationship between knowledge, theory, and performance test scores. An examination of this table shows that the correlations between performance and job knowledge test scores are generally somewhat higher than the correlations between performance and theory test scores; however, the correlations are not high enough to justify the substitutions of job knowledge tests for job performance tests.

Table 1

Correlations of Job Performance Tests With Theory Tests and Job Knowledge Tests

Researchers	Type of Job Performance Test	Theory Test	Job Knowledge Tests
Evans & Smith (1953)	Troubleshooting	.24 & .36	.12 & .10
Saupe (1955)	Troubleshooting	-	.55
Brown, Zaynor, Bernstein, & Shoemaker (1959)	Test Equipment	-	.29
	Alignment	-	.28
	Repair Skills	-	.19
Williams & Whitmore (1959)	Acquisition Radar Maintenance (Inexperienced Subjects)	.03	.36
	(Experienced Subjects)	.14	.22
	Target Tracking Radar Maintenance (Inexperienced Subjects)	.24	.33
	(Experienced Subjects)	.20	.38
	Missile Tracking Radar (Inexperienced Subjects)	.09	.15
	(Experienced Subjects)	.19	.32
	Computer (Inexperienced Subjects	.08	.24
	(Experienced Subjects)	.06	.14
Crowder, Morrison, & Demaree (1954)	Troubleshooting	.11	.18 & .32

Note: From Foley, 1974.

Foley goes on to discuss paper-and-pencil simulation of the troubleshooting task through the use of "Tab Test" techniques in which the subject is provided with an equipment schematic and a description of the initial indications that a problem exists. The subject obtains test point readings and other information by lifting tabs, erasing a covering layer, etc. as he goes through a written description of the troubleshooting process. Crowder, Morrison, and Demaree (1954) reported correlations of .12 and .16 between two forms of a Tab Test and a performance test on the actual equipment. Steinemann (1966) found correlations that ranged from -.50 to +.14 between various measures yielded by a Tab Test designed to measure ability to troubleshoot a superhetrodyne receiver and comparable measures from a test on the actual equipment. Steinemann points out that, in the simulated troubleshooting test, measures are obtained by simply erasing the covering material. Consequently, the measure given is always accurate. In actual practice, checks and measurements require considerably more effort and the accuracy of the reading depends upon the skill with which the test equipment is used. Steinemann found that, when taking the actual performance test, subjects often repeated the same check or measure because they were uncertain of the accuracy of their findings. He states:

> Dubious measurements tended to affect the entire troubleshooting sequence. Reliance upon an incorrect reading, for example, could lead examinees to a false casualty assumption. Conversely, uncertainty over a correct reading sometimes caused students to persist in repeating an unproductive line of troubleshooting strategy . . . In the actual task, students were reluctant to unsolder or disconnect components from the chasssis, but in the simulated task, where parts replacement required virtually no effort, students too often resorted to parts replacement in an effort to solve the problem (pp. 10-11).

Steinemann concluded that the evidence strongly suggests that caution should be exercised in assuming that any simulated performance measure, even when it has considerable common identity to the actual task, will provide a valid estimate of proficiency on the actual equipment:

> In the development of Navy-wide systems for proficiency evaluation, simulated performance measures should be empirically validated against real performance criteria, before they are accepted for incorporation into the total assessment system (p. 11).

From 1969 to 1975, the Advanced Systems Division of the Air Force Human Resources Laboratory supported "a modest program to provide the Air Force with the necessary tools for measuring the ability of maintenance personnel to perform the key tasks of their jobs" (Foley, 1975, p. 4). The objectives of this program were not only to develop a model battery of performance tests on electronic maintenance tasks but also, using the model battery as criteria, to develop and try out a series of paper-and-pencil symbolic substitute tests.

A performance test battery was developed using the AN/APN-147 doppler radar and its computer, the AN/ASN-35, as test vehicles. The test was designed to cover the various types of organizational and intermediate maintenance activities which Air Force electronic technicians engage in; thus, tests were developed on troubleshooting, alignment/adjustment/calibration, use of test equipment, operational checkout, soldering, and component removal/replacement. Scoring was on a "go, no-go basis in terms of meeting job product specifications" (Shriver & Foley, 1974).

A separate paper-and-pencil "graphic symbolic substitute" test was developed for each performance test. In developing these tests, it was hypothesized that symbolic substitute tests could be developed that would be more valid than previously developed symbolic tests because they would contain more realistic task "clutter." For example, in testing troubleshooting, if the subject wanted to know the voltage at a specific test point, he would be provided with a picture of a voltmeter that displayed the requested information in the same way he would see it on the actual equipment, rather than being provided with a printed voltage readout. The symbolic tests were subjected to a small-scale validation in which novice technicians took both the symbolic and performance versions of the tests. On the basis of these results, the symbolic troubleshooting test was modified and then subjected to validation utilizing experienced technicians. The results of both validations are summarized in Table 2. It was concluded that the symbolic tests, with the exception of the one on soldering, showed sufficient promise to justify further consideration and refinement. (It should be noted that the soldering task has the largest motor-skill component and, consequently, it would be least likely to be amenable to symbolic testing.) The authors point out that valid symbolic substitute tests cannot be developed for any job activity until good job-performance tests are available (Shriver & Foley, 1974).

Vineberg and Taylor (1972a, 1972b) examined the relationship between job sample and job knowledge test scores in four Army jobs (i.e., armor repairman, vehicle repairman, supply specialist, and cook). In all four of these jobs, the skill component, in contrast to the knowledge component, was judged to be minimal. Consequently, it was hypothesized that job knowledge tests could be appropriately substituted for job sample tests. In contrast to Foley and Shriver's work, where the stress was on paper-and-pencil simulation of essential task elements, the stress in this work was on multiple choice, job-related knowledge items. Table 3 shows the correlations that were obtained when both job sample and job knowledge tests were administered to personnel in each of these job areas. The authors indicate that these results tend to support their hypothesis that "job knowledge tests can be appropriately substituted for job sample tests, when a job contains little or no skill components and when only knowledge required on the job is used on the test" (Vineberg & Taylor, 1972b, p. 19). It should be pointed out, however, that while the data lend support to the authors' general hypothesis the relationships are far from perfect. It does not appear that these knowledge tests could be utilized as adequate substitutes for the performance tests in a situation in which fairly precise information was sought concerning job-skill strengths and deficiencies. The knowledge tests which were constructed were based upon task analysis information and they were designed to contain only information which was clearly relevant to job performance; however, they were not completely parallel to the performance tests in terms of content and level of knowledge required. It would appear that, in these four areas, the degree of relationship between job knowledge and job performance tests could be raised through further refinement of the knowledge tests. Whether or not these knowledge tests could be improved to the point where they could be confidently substituted for performance tests remains an open question.

Table 2

Summary of Relationship Found Between Performance and Symbolic Versions of Electronic Maintenance Tests

Test Area	Number	Phi Coefficient	Tetrachoric Correlation
Novice Subjects			
Checkout	4	1.00	-
Removal and Replacement	14	.43	-
Soldering	4	0	-
General Test Equipment	6	.67	-
Special Test Equipment	6	.33	-
Alignment and Adjustment	19	.58	-
Troubleshooting	9	-.33	-
Experienced Subjects			
Overall Troubleshooting	30	.47	.68
Chassis Isolation	30	.73	.81
Stage Isolation	30	.33	.46
Piece/Part Isolation	15	.07	.16

Note: From Shriver and Foley, 1974.

Table 3

Correlations Between Job Sample and Job Knowledges Tests for Four Army Jobs

Job	Zero-Order Correlations	Partial Correlations[a]
Armor Crewman (N = 368)	.68	.49
Repairman (N = 360)	.59	.49
Supply Specialist (N = 380)	.72	.65
Cook (N = 366)	.58	.50

[a]Correlations with the effects of time on the job partialled out.

Note: From Vineberg and Taylor, 1972(b).

Process vs. Product Evaluation

All job tasks involve both a product or outcome and a process or procedure. A decision which must be made early in the development of a job performance test is whether the test should concentrate on the task process, the task product, or some combination of the two.

In a Navy Job-Performance Proficiency Assessment System, performance test scores would be used for two primary purposes: (1) to provide measures of Fleet readiness in terms of ability of individuals and teams to carry out critical skills and (2) to provide diagnostic information to personnel decision makers concerning the specific reasons why personnel can not carry out assigned critical tasks. Typically, product measures provide information relative to the question of personnel readiness, and process measures provide diagnostic information relative to the question of reasons for deficiencies. Because the proposed assessment system is concerned with both readiness measurement and diagnosis, it can be assumed that, in general, when performance tests are used they will be concerned with both process and product measurement. However, when testing Fleet personnel on critical aspects of their jobs, this may be more easily said than done.

In 1962, Wilson indicated that in performance testing there is a tendency to utilize product measures in place of systematic detailed measurement of specific task characteristics. This situation appears to have changed very little over the years.

Osborn (1974, p. 2) points out that, when dealing with the question of process versus product, any task can be considered to fall into one of these three types:

1. Those in which the product is the process.

2. Those in which the product always follows from the process.

3. Those in which the product sometimes follows from the process.

In the military, there are very few tasks of the first type; that is, those in which the task and process cannot be easily separated. Normally, these would be tasks that serve mainly an aesthetic purpose, such as gymnastics and figure skating. A military example is close-order drill.

A great many military tasks are of the second type; that is, fixed-procedure tasks in which the product always follows from the process. Military examples are use of electronic test equipment and the performance of many maintenance checks. For this type of task, "the procedural steps are known, observable, and comprise the necessary and sufficient conditions for task outcome; so, if process is correctly executed, task product necessarily follows" (Osborn, 1974, p. 2).

The majority of military tasks are of the third type in which the product does not always follow from the process as we are able to measure it. This can be because (1) we are unable to specify all of the steps required to

perform a task, (2) there are a number of paths that can be followed in order to arrive at a correct solution and we are unable to specify all of these, or (3) we are unable to measure accurately one or more procedural steps.

The sonar operator's task of classifying sonar contacts, particularly with active sonars, is a good example of this third type of task. If all of the audio, video, and environmental cues which are used by sonarmen to arrive at classification decisions could be specified and reliably measured and the ways in which sonarmen might put these cues together was completely understood, the classification task would be of the second type and the process could be measured without concern for measuring the product. However, this is not the case. We simply cannot specify and reliably measure all possible cues that a sonarman might use and all possible ways he might put those cues together to arrive at a correct classification decision. Consequently, evaluation of sonar classification skills tends to concentrate upon the product, which is the classification decision.

Osborn (1974) makes the point that "Because of the interchangeability of process and products for tasks of the first two types, it does not really matter which measure is used to assess proficiency; but for tasks of Type 3, product measurement is very important" (p. 3). It should also be emphasized that if both proficiency levels and the reasons for detected deficiencies are of concern, then both process and product measures are important.

Highland (1955) has provided a summary of factors that the test developer should consider when attempting to determine the relative weights to be given to process and product measures.

> The following conditions will make it more likely that the test constructor will wish to score performance in terms of the process--that is, in terms of how something is done.
>
> 1. The steps in a procedure can be specified and have been explicitly taught.
>
> 2. The extent to which an individual deviates from accepted procedures can be accurately and objectively measured.
>
> 3. Much of the evidence needed to evaluate performance is to be found in the way that performance is carried out and/or little or none of the evidence needed to evaluate performance is present at the end of performance.
>
> 4. An ample number of persons are available to observe, record, and score the procedures used during performance.

The following conditions will make it more likely that the test constructor will wish to score performance in terms of products evident after the performance has been completed, and available even though the performance has not been observed.

1. The product of performance can be measured accurately and objectively.

2. Much of the evidence needed to evaluate performance is to be found in the product available at the end of performance and/or little or none of the evidence needed to evaluate performance is to be found in the way that performance is carried out.

3. The proper sequence of steps to be followed in attaining the goal is indeterminate, or has not been taught during training or when, though everyone knows the steps, they are hard to perform and skill is ascertainable only in the product.

4. The evaluation of the procedures used during performance is not practicable because persons are not available to observe, record, and score these procedures. (p. 34)

Schmidt, Greenthal, Berner, Hunter, and Williams (1974) made the following points in favor of end-product evaluation:

1. It may be less difficult and time consuming to train non-psychologists to evaluate final products than to teach them to observe and record specific behaviors.

2. If care is taken in the measurement process, agreement between judges as to the quality of final product is generally at least somewhat higher than agreement as to suitability of specific behaviors (Stuit, 1947; Bornstein, Jensen, & Dunn, 1954).

3. Examinees can be expected to feel less threatened and nervous, since they are not watched as they work. This factor should make performances more representative of the examinee's actual skill level, especially in the case of complex tasks requiring reasoning and problem solving.

4. Evaluation of the end products can be carried out after test administration, at the convenience of the evaluator. He is thus freed to concentrate on test administration.

5. The resulting scores should be more valid. The primary concern in real life is the ability to produce high quality end products. Specific behaviors, tool usages, task sequences, etc., are merely means to an end and are therefore of secondary concern. (pp. 15 & 16)

While describing procedures which should be followed when developing occupational competency tests, Panitz and Olivo (1971a) indicated the following:

> Work sample selected for the performance test must permit an evaluation of work habits as well as permit an objective evaluation of the finished product. The choice of the work sample is further influenced by the length of time required for its completion, the amount of equipment required and accessible, availability of testing personnel and testing facilities, funds, etc. . . . In any case, both method and product must be evaluated. (p. 33)

When both product and process measurements are obtained, they typically are lumped together so that a total score can be provided. Such total scores have little relationship to the job situation and are extremely difficult, if not impossible, to interpret. In a Navy Job-Performance Proficiency Assessment System, test results would probably be analyzed to provide information concerning both proficiency levels (products) and specific deficiencies.

In some situations, it may be possible to measure products and then revert to process measuring when deficiencies are detected at the product level. Such an approach, however, would have to be utilized with considerable caution because for many tasks there are critical steps where failure to perform will not always, or even usually, show up in the end product. For example, failure to observe safety precautions when troubleshooting electronic equipment will usually have little effect on whether or not a malfunction is found; however, it may have a marked effect on the rapidity with which an electronic technician leaves the Navy.

Objectivity

When performance tests are utilized, it appears that there is an almost universal assumption that, because such tests involve real-world activities, they will automatically be valid and reliable. In addition, it is assumed that, with minimal training, test administrators will be able to evaluate performance objectively. When developing performance tests for use in a Navy Job-Performance Assessment System, no such assumptions should be made. Every effort should be made to assure that tests are valid, reliable, and objective.

Objectivity as used here pertains to the consistency with which examiners make their judgments. A test is objective when different examiners can use it to observe the same individuals at the same time and obtain comparable

results. Another indication of objectivity would be when the same examiner can use the test to make the same, or nearly the same, observations when presented with identical situations at different points in time.

When performance testing programs are developed, steps are seldom taken to assure that the objectivity of the test instruments will be maximized. Shimberg, Esser, and Kruger (1972) point out that the most serious deficiency in performance tests used in occupational licensing examinations is "the lack of adequate criteria or standards for evaluating performance." They feel that specific directions as to what they are to look for, what constitutes acceptable performance on a given task, and how much credit should be deducted for failure to satisfy the criteria in specified ways. Without such guidelines, they are forced to use subjective measure that are based upon their own experience and standards (p. 198).

Siegel (1954) described the ideal method of determining objectivity as one in which the examinee's performance is held constant over two separate occasions and the observer's judgments are allowed to vary. He had film sequences made on naval aviation structural mechanics as they took a drill point grinding work sample performance test. These film sequences were shown twice to five observers with a 1-month interval between showings. While they viewed the film, each observer filled out a checklist concerning specific steps in the grinding process. Objectivity (i.e., intraexaminer consistency) was determined by dividing the number of steps evaluated in exactly the same manner on each showing by the total number of evaluations that the checklist required. The percentages of times each observer made consistent observations were 64, 71, 86, 92, and 100 with a mean of 83.

In developing a diagnostic test equipment test, A. Abrams (1962) compared the performance of two raters over a series of 204 observations. Agreement occurred on 201 of these observations, a percentage of agreement of 98.5. This high degree of agreement appeared to be a function of both the care that went into the development of the test and the training which was given the raters.

Schmidt et al. (1974) developed a set of performance tests for metal trades apprentices (i.e., horizontal and vertical mill, drill press, lathe, and surface grinder), which were designed for administration by personnel with little or no metal trade experience. The researcher reported a high degree of agreement between laymen on the evaluation of end products. Apparently, this was achieved by paying close attention to the development of scoring techniques that minimized the subjectivity of the evaluation process. Tolerence evaluations were made with the aid of dial-read precision instruments, and finish evaluations were made with the aid of "benchmarks," i.e., finished products chosen in advance to correspond to specific quality levels of finish. The authors concluded that the findings not only reflected favorably on the scoring technology used, but also demonstrated that laymen with no previous experience in the metal trade or in using precision instruments can quickly be taught to evaluate end products. It should be pointed out, however, that all of these tasks resulted in end products that were relatively easy to evaluate in terms of specific finish requirements.

The U.S. Army is currently concerned with the development of job performance tests for use in its annual MOS testing program (McCluskey, Trepagnier, Cleary, & Tripp, 1975a; 1975b). Prototype performance tests were developed that covered the tasks of (1) preparing a M72A2 light antitank weapon (LAW) for firing and restoring it to carrying configuration, (2) installing and recovering an electrically armed Claymore mine, (3) applying life-saving measures, and (4) camouflaging self and individual weapon. These tests were subjected to a field evaluation which involved five groups. Each group consisted of 15 examinees, 1 test administrator who was responsible for site preparation and administration of all tests, and 4 observers who were required to independently rate each examinee as he performed each test. On all tests the agreement between raters was found to be quite low. The following factors were considered to be the main sources of this disagreement:

1. The evaluation of several performance measures was dependent on the examinee's verbal report, which may have created a situation of low reliability because of different terminology or audibility.

2. Some performance measures required that the preceding actions be completed in a specific sequence, and it appeared that some raters strictly followed the criterion while others did not.

3. Some performance measures were ambiguous statements which were open to the interpretation and bias of the individual rater.

4. Some performance measures appeared to be interpreted differently as a function of specific unit SOPs.

5. Two separate actions were included in single performance measures, and it is not known whether the raters required one or both of the actions to be completed for a "Yes" score.

6. Several performance measures overlapped or were repeated later in the sequence which appeared to create confusion in scoring among the raters.

7. Sequences in the scenario which were timed were not clearly identified, and this appeared to result in different evaluations of whether or not the time standard had been met.

8. Some performance measures in the scenario were out of sequence with respect to the order in which they would be scored. (McCluskey et al. 1975b, p. 32)

It was concluded that the objectivity of the tests could be raised to satisfactory levels by improving the test development process and providing additional instruction and training for the raters. However, even if a satisfactory level of objectivity can be achieved during initial test development, it remains to be determined whether a high degree of objectivity can be maintained in an operational setting over a long period of time.

An interesting aspect of this study was an evaluation of the degree to which the various administrators followed the prescribed administration procedures. The researchers found "numerous examples" of the introduction of individual interpretation and bias into the test situation. These findings are summarized in Table 4. In approximately 37 percent of the administrations, at least one deviation was observed from the prescribed testing procedure. The researchers indicated that the NCOs were given specific instructions not to coach the examinees or to provide any feedback during the test; however, these NCOs found it very difficult to divorce themselves from their normal training role. This same phenomenon has been observed frequently when this Center has utilized military job experts to administer performance tests. In a performance testing system designed to obtain information over a long period of time, the problem of maintaining standardized test administration procedures would appear to be a serious one. Various solutions to this problem should be investigated. One such solution might be to present standardized sets of instruction by means of either closed-circuit television or motion pictures.

Osborn (1975) has suggested that objectivity of performance test scoring can be increased by combining scoring templates with films (video tapes) of critical performance sequences:

> Where the model response on a test of marksmanship is defined as a hole in the bullseye, it is relatively easy for the scorer to judge the acceptability of the response made by the rifleman. The concentric circles normally marked on a target act as a kind of simple template which enhances the ease and objectivity of scorer judgments. Templates could be applied equally well in scoring other tests. For example, tasks . . . in which the outcome is a process are often difficult to assess reliably. It would appear that performances such as springboard diving or gymnastic exercises could be more objectively scored if the outcomes were filmed and figural templates overlayed on key frames to assess the performer's accuracy at those critical points. . . . For these particular tasks--or for that matter, any task in which the product is transient--the added cost in recording the product for later scoring would probably be offset by savings in scoring costs; that is, the more objective approach to scoring would very likely preclude the usual requirements for a panel of expert evaluators. But more important, the scorer would not be constrained by real time, and could function at a place and time and rate of his or her choosing, using prepared templates to increase objectivity. (pp. 89-90)

Table 4

Variability in Test Administration

Characteristics of Test Administration	Percent of Administrations Given "Yes" (Total N = 75 for each test)			
	LAW	Camouflage	Life Saving	Claymore
1. Instructions were read verbatim to the examinee from the test scenario	75.7	79.2	73.5	43.3
2. All activities described in the test scenario were completed	61.6	56.9	60.3	71.6
3. The test administrator coached or provided feedback to the examinee during the test	35.6	26.4	57.4	19.4

Note: From McCluskey, Trepagnier, Cleary, and Tripp, 1975b.

Along these same lines, Boyd and Shimberg (1971) indicate that video recorders could be used to train observers. After some initial training on the administration of a specific test, prospective observers could view a video tape of an examinee taking the test and rate the examinee's performance. Then, any discrepancies in the ratings could be discussed and resolved. After that, the group could be asked to rate additional videotaped examinees. Observers who continue to make divergent ratings might either be exposed to further training or eliminated from the testing program. Also, Boyd and Shimberg indicate that "in some situations where process evaluation is of great importance, it may be possible to record an entire test performance on videotape so that two or more observers could rate the individual after the performance has been completed. Where ratings differed, judges could re-examine the tape together, discuss the behavior in question, and resolve their differences; or a neutral judge could be called in to assist in reaching a decision" (p. 29).

Reliability

Cogan and Lyons (1972), when discussing the use of performance tests for quality control of training programs make the point that the term reliability has multiple meanings and use of the term is more likely to confuse than to clarify. A committee of measurement experts recognized this when they stated that: "Reliability is a generic term referring to many types of evidence. The several types of reliability coefficients do not answer the same questions and should be carefully distinguished" (Technical recommendations for psychological tests, 1954, p. 28). This committee recommended that a distinction be made between: (1) a coefficient of internal consistency based on

internal analysis of data obtained on a single trial of a test, (2) a coefficient of equivalence based on the correlation between scores from two forms given at essentially the same time, and (3) a coefficient of stability based upon the correlation between a test and retest, with an intervening period of time.

Most studies that deal with the problem of the reliability of performance tests are concerned with the objectivity of the measuring instruments. There is a scarcity of studies which deal with either the internal consistency, equivalence, or stability of performance tests. Schmidt et al. (1974) reviewed the research literature on performance testing for the period from 1922 to 1972. They found only three studies that contained internal consistency coefficients and only one with a test-retest coefficient. The picture has changed very little since that time. Seventeen studies, published since the review, were identified in which performance tests were utilized as data gathering instruments. Fifteen contained no reliability information, one contained data concerning objectivity of the measurement instruments, and one contained both objectivity data and a split-half coefficient of internal consistency.

Validity

As with reliability, the term validity has many meanings. Consequently, it is important to indicate clearly what the term means in the job performance test context. In general, the degree to which a job performance test is capable of achieving its aim is evaluated in terms of either content or concurrent validity.

It is generally agreed that a job performance test can be considered to be valid if, for the skill in question, it discriminates either between those who are proficient and those who are not proficient or among individuals with various degrees of the skill (DuBois, Teel, & Petersen, 1954; Glaser & Klaus, 1962). When the validity of a performance test is evaluated by means of such a comparison, it is referred to as concurrent validity.

In many situations when job performance tests are being validated, no adequate independent method is available for distinguishing between individuals with various degrees of the skill in question. Consequently, concurrent validity estimates cannot be obtained. In these situations, the content validity of the test is evaluated. This requires a determination of the degree to which an individual's responses to the test situations can be considered to be a representative sample of the responses that would be obtained from the universe of situations that constitute the area of concern. In most cases, it is not possible to obtain quantitative evidence of content validity, and the argument for such validity is established deductively by defining a universe and sampling from that universe (Lennon, 1956; Technical recommendations for psychological tests, 1954).

Cronbach (1969) points out the following:

> Content validity is necessarily limited by the inadequacy of the universe specification, which is usually couched in imprecise,

> everyday terms and can rarely mention every pertinent aspect of the task. Content is an ill-shaped and undifferentiated mass, hence there is a danger of vagueness in any reference to a content universe. Moreover, while there may be a definable domain of content, there is no existing universe of items. The only items in existence are likely to be those that constitute the so-called sample (p. 43).

The central requirement is that the universe boundaries be well defined. This requirement presents a very real challenge when one attempts to develop performance tests which cover critical aspects of complex jobs.

Cronbach (1969) further points out that there is nothing in the logic of content validation which requires that the content of a test be homogeneous. To make a decision concerning automobile license applicants, it is necessary to know whether the drivers can perform various tasks defined by the universe of good driving practices. If the tasks have low correlations, it will take a larger sample of tasks to be confident that the subject is at the appropriate proficiency level. Cronbach states, "No matter how heterogeneous the universe, with enough items one can estimate the universe score as precisely as desired. Low item correlations do not necessarily imply failure of the test content to fit the definition. Indeed, if the universe is heterogeneous, consistently high correlations imply inadequate sampling" (p. 45).

He goes on to mention that correlations between tests are irrelevant in terms of content validity. Thus, even where high correlations exist between knowledge and performance tests, there is strong justification for keeping the two measures separate. First, the level of attainment might be much higher for one than the other and this could have implications for program modification. Second, although the two tests correlate at one point in time, circumstances could change (e.g., curriculum modifications, changes in job content), which would result in a drastic change in the relationship between the two tests.

A factor that must be considered when discussing the validity of job performance tests is that such tests do not provide a final measure of how an individual will perform on the job. Under the actual job situation, an individual "may have to perform these tasks in cramped quarters; under stress of time, noise, heat, or cold; or with an excited boss interfering. These conditions of stress are usually not constant variables, but change from day to day and hour to hour" (Foley, 1974, p. 10). Under these conditions, it usually has to be assumed that the individual can perform a task under conditions of stress, if he can perform the same task well under normal conditions. Conversely, it is assumed that, if an individual cannot perform a task in the absence of stress, he will not be able to perform that task when a variety of stresses is present. It should be pointed out, however, that, with some individuals and/or under some conditions, the testing situation may actually be more stressful than the job situation.

When one looks at what happens in actual practice, it becomes obvious that very little attention is paid to the question of the validity of performance tests. In most cases, it is apparently assumed that, when one develops performance tests, they are automatically valid and, of course, they are also automatically objective and reliable. However, there are a few research studies that have concerned themselves with the problem of validating performance tests.

Popham (1971) attempted to validate a test of teacher effectiveness by testing the hypothesis that "the performance test at least ought to be able to discriminate between experienced teachers and nonteachers with respect to their ability to accomplish prespecified instructional objectives" (p. 109). In this performance testing approach, an instructor was given a set of objectives and all necessary resource materials, and he was then told to prepare an instructional sequence that would accomplish the objectives. After a suitable period of time, he was required to teach to the objectives using whatever instructional procedures he had developed. A test of student performance was used to evaluate the degree to which the instructor had taught the specified objectives. Validations of this performance testing procedure were carried out in three areas: (1) social science research methods, (2) basic electronic power supplies, and (3) automobile carburation. For the social science objectives, the performance of 13 experienced teachers was compared to the performance of 13 college social science majors. For the power supplies objectives, the performance of 28 experienced teachers was compared to that of 28 experienced electronic technicians. For the carburation objectives, the performance of 16 experienced teachers was compared to the performance of 16 experienced automobile mechanics. The final results showed that "in all three instances there were no significant differences between the ability of teachers and nonteachers to promote learner attainment of prespecified objectives" (Popham, 1971, p. 113). Popham indicated that he believed the measuring instruments were quite satisfactory and the results obtained were probably brought about because "Experienced teachers are not particularly skilled at bringing about prespecified changes in learners" (p. 115). On this task, experienced teachers were simply no better qualified than inexperienced teachers. This research illustrates a problem that can exist when one attempts to validate a performance test on the basis of its ability to discriminate between experienced and inexperienced groups. As observed many times in performance testing carried out by this Center, experience does not necessarily result in an individual who is skilled at performing his job duties. Practice does not always make perfect. A good example of this is automobile driving, where it would not be at all surprising to find inexperienced drivers who had just participated in a driving training program to have skills which are superior to those of experienced drivers.

Jones and Whittaker (1973) encountered this problem of interpreting the relationship between experience and skill level when they attempted to validate a performance test for factory workers. This test was designed to measure the proficiency of trainees after (1) a bench-fitting course for apprentices and (2) a short electronics course for apprentices and older personnel. The validation involved comparing the performance of personnel who had just completed the courses with that of personnel who had completed the courses two years previously and, thus, had 2 years of on-the-job experience. On the fitting skills test, the experienced group was superior to the inexperienced group on two of the six tests--selecting and using reading instruments and completing the marked-out job. There was no significant difference between the two groups on interpreting technical drawings, marking out the job, planning marking out, and planning metal removal. On the electronics test, the inexperienced group was superior on four of the eight tests--recognizing color codes on transistors, identifying circuit test points, using test

instruments, and finding faults. There was no significant difference on recognizing circuit symbols, choosing test instruments, recognizing diagrams of subcircuits, and recognizing color codes on resistors. The authors speculated that the superior performance of the experienced group on two of the fitting tests was the result of practice on the job and the lack of any differences on the other tests showed that the skills being measured were not often practiced on the job. Furthermore, they indicated that the superior performance of the inexperienced electronics group on fault finding, the key objective of the course, and on certain other skills raised questions about the validity of the electronics test battery. They noted, however, that it was probable that the performance patterns that were obtained were the result of the skills not being utilized on the job.

Whipple et al. (1969) validated a test of the job effectiveness of Army radar mechanics by (1) comparing the performance scores of mechanics at the time of their graduation with the scores of those with varying amounts of field experience and (2) correlating performance test scores and school grades. The correlations obtained ranged from .53 to .60. As shown in Figure 1, it was found that performance scores increased with experience for two of the subtests but not for the third.

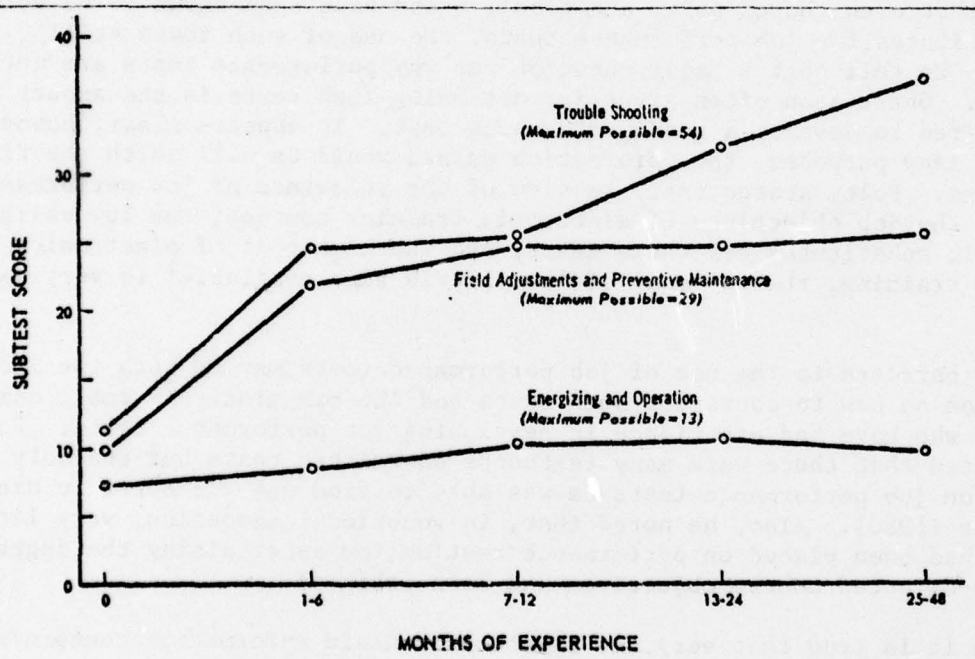

Figure 1. Subtest performance of groups differing in experience (from Whipple et al., 1969)

These results were interpreted as providing strong evidence for the validity of the test. If the authors had been concerned about the validity of the subtests rather than the overall test, it would probably have been concluded that the troubleshooting and field adjustment subtests were valid but that the validity of the energizing and operation test was questionable.

It was mentioned previously that, in general, when performance tests are used, their validity is taken for granted. To illustrate this point, a search of the literature was conducted in which 34 reports were identified that involved use of specially developed performance tests to measure the proficiency of various types of individuals. In 27 of these reports, no mention of validity was found, one indicated the test had high empirical validity by definition, four contained information on concurrent validity, and two contained information related to content validity. Illustrative of what typically happens is a project that was undertaken to develop and validate occupational examinations for seven occupational areas (i.e., automobile mechanics, industrial electronics, machine trades, machine drafting, residential electrician, chef, and carpentry) (Ross, 1973). Both written and performance tests were developed. The written tests were validated, but not the developed performance examinations.

Utilization of Performance Tests

Foley (1974), when discussing the evaluation of electronics technicians, observed that even though paper-and-pencil tests have been shown to be very poor substitutes for job performance tests, the use of such tests still persists. He felt that a major question was why performance tests are not used more. One reason often given for not using such tests is the amount of time required to develop a good performance test. It appears clear, however, that, for many purposes, the information gained would be well worth the time and expense. Foley stated that, in view of the relevance of job performance tests "to the job objectives of electronic training courses, the low validity of symbolic substitutes for these tests, and the high cost of electronics technical training, the rational of 'too little time available' is very weak" (p. 17).

Other barriers to the use of job performance tests may be both the lack of information on how to construct such tests and the comparatively small number of people who have had experience in developing job performance tests. Foley (1974) noted that there were many textbooks on written tests but the only text material on job performance tests he was able to find was a chapter in Michaels and Karnes (1950). Also, he noted that, in vocational education, very little emphasis had been placed on performance testing for ascertaining the degree to which job-oriented course objectives had been achieved.

While it is true that very few textbooks contain information concerning the development of job performance tests, a number of manuals have been prepared which are primarily designed to provide guidelines for use when developing such tests (Boyd & Shimberg, 1971; Highland, 1955; Panitz & Olivo, 1971b; Swezey & Pearlstein, 1974; Vineberg & Taylor, in press; Wilson, Mackie, Buckner, Siegel, & Courtney, undated). Also, the Interservice Procedures for Instructional Systems Development (NAVEDTRA 106A) has an excellent chapter that

describes procedures to be followed in constructing job performance tests. Utilization of these procedures should result in more frequent and better use of performance tests in military schools. It remains an open question whether or not by following such manuals individuals can develop job performance tests that are objective, reliable, and valid.

Panitz and Olivo (1970) investigated "the methods and procedures employed by industrial firms for the evaluation of occupational competency" (p. 36). Their review failed to turn up any published studies on performance testing. They then sought assistance from the American Management Association, the National Association of Manufacturers, the Industrial Conference Board, the American Association of Industrial Management, the National Metal Trades Association, the Associated Electrical Industries, the National Association of Industrial Training Directors, and the National Tool, Die, and the Precision Machine Association. On the basis of leads provided by these organizations, only two companies were identified that carried on a formal program of occupational competency testing. One of these administered a written proficiency examination while the other carried out "simulated performance tests based on a thorough analysis of the tasks in a particular job" (Panitz & Olivo, 1970, p. 37). The authors also reported that the National Industrial Conference Board surveyed 384 firms and found no evidence that any of them carried on any form of practical performance testing. On the basis of these observations of what takes place in industry, Panitz and Olivo (1970) state that:

> There are many reasons why there is so little work published or, in fact, actually done in occupational competency testing in industry. Possibly, industry believes it satisfactory and has found it expedient to follow the practice of closely observing the performance of new employees during their beginning days of employment as the equivalent to a performance test. In the absence of standardized, valid instruments, which may be secured commercially, many medium and small-sized companies have stated that lack of qualified competent personnel, and the cost of developing accurate and reliable instruments, makes the operation of a performance testing program prohibitive. Whether action to not carry on an occupational training program is justified on such grounds is a decision industry must make.
>
> However, the project staff who have spent their industrial and professional lives working in and with industries, in expressing their personal beliefs, feel that valid occupational competency programs should be widely established across all industries, in the labor movement and throughout government agencies. (p. 38)

In surveying occupational testing programs in the military, Panitz and Olivo observed that relatively few papers had been published that were concerned with performance testing and that most of the tests used were of the pencil-and-paper type. They did find, however, that the Army Signal Corps Schools and the Navy Electronics School carried out performance testing programs; consequently, they made special visits to these schools. They observed that the schools visited were organized so that a special group of

personnel was responsible for development, administration, and control of performance tests. They indicated that these programs were the most promising of any they were able to identify. However, there was a considerable variation in the methods and procedures which were followed by the various schools.

In summing up their state-of-the-art study, Panitz and Olivo (1970) indicated that, among other things, the study provided crystal clear evidence:

> That not a single professional testing agency, governmental organization, educational body, labor or management group was satisfied with either the qualitative or quantitative aspects of occupational competency testing;
>
> That an exeedingly limited amount of data or experiences relating to occupational competency development, administration, validation, or research results are recorded;
>
> That in vocational education a Consortium of States effort to develop, administer, and validate tests requires the investment of significant sums for, at least, twenty (20) major industrial occupations, the employment of competent staff on a full-time basis, supportive "software" and "hardware," and a cooperative arrangement for feedback upon which judgments may be made on the effectiveness of each instrument and the value of test results and testing program. (pp. 48-50)

Because of the need they perceived, Panitz and Olivo (1971a, 1971b) went on to develop plans for a National Occupational Competency Testing Institute (NOCTI). This institute was formed in June of 1973 as a nonprofit corporation located at the Educational Testing Service in Princeton, New Jersey. Area Test Centers are being set up throughout the United States. The general purpose of this program is to develop and administer tests that will assure that prospective vocational teachers possess the level of competency that is required for effective teaching in various vocational programs (Panitz & Olivo, 1971b). Thus far, written and performance tests have been developed for 24 occupations (National Occupational Competency Testing--Bulletin, 1974).

Shimberg et al. (1972) carried out an extensive review of occupational licensing practices in the United States. This review covered health occupations, construction trades, service occupations, and transportation occupations. They observed that although many licensing boards recognized the value of using performance tests in judging an applicant's competency, the results obtained from practical testing may be unreliable or misleading because of what the tests include, the way the test is given, or the way that performance is evaluated. A major deficiency in the performance testing carried out by licensing boards was failure to do an adequate job of sampling crucial skills. Some licensing boards were "inclined to select very difficult or even esoteric tasks rather than those that are relevant to demonstrating skills and knowledge" (Shimberg et al., 1972, p. 198). For example, plumbing boards continue to

place undue emphasis on joining lead pipe by means of a lead pipe joint. This skill was once a mark of a true craftsman, but the replacement of lead pipe by copper and the use of prefabricated lead joints has vastly diminished its value in actual job situations. Yet, this task continues to be the only performance task required by a number of plumbing licensing boards. Another deficiency found by Shimberg et al. (1972) was a failure of licensing boards to develop standardized procedures to be followed in the administration of tests. However, in the authors' opinion, the most serious deficiency of performance tests used by licensing boards was a lack of adequate standards for evaluating performance. Raters were not always provided with clear, specific directions as to what they were to look for, what constituted acceptable performance, how much credit was to be deducted for failure to perform in a specified way, etc.

As mentioned earlier, the Army is concerned with the development of job performance tests as part of its annual MOS testing program (McCluskey et al., 1975a, 1975b). These new tests are called Skill Qualification Tests and they are intended to be used for both personnel management (promotion, retention, assignment) and training feedback purposes (Maier, 1976). One of the requirements set for such tests was that they must be fair and feasible, fairness meaning that every soldier would have an equal opportunity to demonstrate his competence. This meant that testing conditions would have to be sufficiently constant throughout the Army so that scores from administrations under different conditions would not be noticeably different. Because the tests are to be administered worldwide, the feasibility requirement meant that tests had to be suitable for administration with all types of units, equipment, terrain, and personnel. Also, all testing materials had to be readily available and testing of a given soldier could not take more than 1 day. It was found that these fairness and feasibility requirements, coupled with the requirement that large numbers of men be tested each year, put severe limitations on the use of job performance tests. Thus far, the resolution has been to have all Skill Qualification Tests contain a written component and some tests contain a job performance component. Four hours of testing time is allowed for the written test and up to 4 hours for the performance test.

A QUALITY CONTROL APPROACH FOR PROFICIENCY ASSESSMENT

From the foregoing it would appear that job performance testing of some sort may provide the best information obtainable on the capabilities of job incumbents. It also appears that performance testing is both difficult to do well and very expensive. Job performance tests are difficult to develop, costly to administer, and may yield results which are difficult to interpret. Clearly, applying job performance testing techniques to everyone in the Navy would be prohibitively expensive.

Fortunately, it is neither necessary nor desirable to administer job performance tests to all Navy personnel with respect to all aspects of job performance. It is not necessary for a number of reasons. First, the capabilities of incumbents in many jobs are self-evident from the performance of those jobs. For example, if a cook prepares a bad meal, it is quickly evident to all diners; if someone dials a telephone incorrectly, he gets a wrong number. Many such instances could be cited. In many other cases, the supervisor's guidance and judgment could be relied upon to ensure adequate performance by the job incumbent. Application of job performance testing to everyone is not desirable since it has been well established that any mass application of measurement processes, particularly when the results are used to determine benefits, leads to its degradation. This has been repeatedly shown relative to the administration of rating scales where almost inevitably the average rating given tends to move toward the high end of the scale with continued use of the instrument. Perhaps the answer to the problem may lie in the use of some modification of quality control techniques developed for the manufacturing of industrial products. Peach (1964), in describing industrial quality control, has said:

> Quality control has usually been associated with quality improvement. It is possible to use QC for quality improvement, but control means, not improvement, but the attainment of management's objectives. If these can be attained by improving quality, good. If their attainment calls for degrading quality, QC can help to do this too. Most often their attainment calls for <u>stabilizing</u> quality, for consistent adherence to a policy chosen by the manager, rather than haphazard and unpredictable fluctuation from good to bad and bad to good again, which is characteristic of all uncontrolled processes. In this job modern statistical quality control can succeed where no other method can, because it not only provides the operational means for implementing a quality policy, but also creates uniquely suitable reports that will tell the manager whether his quality objectives are being attained or not (p. 18).

In a later discussion of sampling inspection, Peach says "We have here an illustration of a basic principle of quality control; that it is more effective to control the <u>process</u> than the <u>product</u>. Once poor quality product has been produced, inspection and salvage can accomplish little, and that only at considerable expense" (p. 124).

OPNAVINST 1500.19C states that:

> A principal objective of the personnel and training organization of the Navy is to provide to the Fleet Commanders in Chief trained personnel capable of maintaining, operating, and employing effectively the ships, aircraft, and weapon systems with which Fleet Commanders in Chief have been equipped and also to provide trained personnel for logistic and other essential support.

This is the operational statement of the requirement placed on the Navy's personnel system. If that requirement is to be carried out effectively and efficiently, it means that the processes that bring individuals to occupy the Navy's jobs must be both effective and efficient. That is, those processes should not only provide fully capable job incumbents but they should not waste resources on providing personnel capabilities unrelated to the requirements of their jobs. It would seem that the quality control approach would be an appropriate one. For the Navy's personnel system, this approach requires:

1. An identification of critical and important tasks.
2. The establishment of appropriate performance criteria relative to critical tasks.
3. The development and implementation of appropriate sampling procedures relative to both tasks and job incumbents.
4. The development and application of effective procedures for measuring the performance of job incumbents in quantifiable terms.
5. An understanding and quantification of the personnel processes that bring individuals to their jobs.
6. A capability for analyzing performance data and relating results to appropriate personnel processing practices.
7. A capability to provide personnel managers with appropriate and understandable reports.

Such a system would not be concerned with evaluating individuals or Navy units. Its purpose would be to supply appropriate personnel managers with information on how well their systems are working. Another quote from Peach is pertinent: "The fact that a product is or is not within tolerences has almost no bearing on quality control. In QC, the maker of the product seeks to achieve mastery of the process, so that he can guide it the way he wants: this is what control means" (1964, p. 141).

In general, what is required is a system which will yield reliable and accurate information on the effectiveness of the personnel processes which bring individuals to their positions in the Navy man-machine system. These processes include recruitment, school training, assignment, on-the-job training, and job experience. This is the kind of information that the manager of one of these processes needs in order to support his decision making. What is needed is a personnel quality control system. The recently developed Interservice Procedures for Instructional Systems Development recognize the need for periodically monitoring the products of military training programs by means of well designed performance tests:

> The fact that today's graduates can do the job they were trained to do does not mean tomorrow's graduates will do the same. The job may change, students may change, something in the course may change, the qualities of instructors may change. Assuring optimum training quality at minimum cost demands a constant feedback of information, and periodic evaluation of the relationships between students, the instructional program, and job performance in the field (Interservice Procedures for Instructional Systems Development, 1975, Vol. V, p. 83).

CONCLUSIONS

Some form of job performance testing appears to be the best source of incumbent capability information relative to critical Navy tasks. It also is very clear that useful job performance testing is difficult, expensive, and demands substantial expertise. It is evident that the costs of measuring all aspects of job performance for all Navy incumbents would be prohibitive. It is also evident that, where job performance tests have been developed and used, typically insufficient attention has been given to evaluating the quality of the performance measurement instruments themselves.

RECOMMENDATIONS

It is recommended that:

1. The Navy develop and evaluate a prototype proficiency assessment system based on the quality control approach suggested in a previous section.

2. A series of experimental studies be conducted to provide additional information and guidance on job measurement techniques relative to such questions as objectivity, reliability, validity, degree of job simulation required, and guidelines for the use of testing procedures in measurement programs.

REFERENCES

Abrams, A. J. Experimental training of sonarmen in the use of electronic test equipment. I. Diagnostic testing of basic sonar students (Technical Bulletin 62-1). Washington, D.C.: Bureau of Naval Personnel, January 1962.

Abrams, A. J., & Klipple, A. G. Accuracy and consistency in judging active sonar classification cues: III. Video cue judgments (NPRA Technical Bulletin 66-17). San Diego: U.S. Naval Personnel Research Activity, December 1965.

Ainsworth, L. L., & Bishop, H. P. The effects of a 48-hour period of sustained field activity on tank crew performance (HumRRO Technical Report 71-16). Arlington, VA: Human Resources Research Organization, July 1971.

Anderson, A. V. Training, utilization, and proficiency of Navy electronic technicians. VII. An overview (Technical Bulletin 63-4). Washington, D.C.: Bureau of Naval Personnel, March 1963.

Anderson, A. V., & Pickering, E. J. An informal investigation of the proficiency of advanced sonarmen in the use of test equipment (Memo Report 59-2). San Diego: Naval Personnel Research Activity, July 1959. (a)

Anderson, A. V., & Pickering, E. J. The proficiency of Pacific Fleet sonarmen in the use of electronic test equipment (Technical Bulletin 59-30). Washington, D.C.: Bureau of Naval Personnel, November 1959. (b)

Bornstein, H., Jensen, B., & Dunn, T. The reliability of scoring in performance testing as a function of the tangibility of the performance product. American Psychologist, 1954, 9, 336-337.

Boyd, J. L., Jr., & Shimberg, B. Handbook of performance testing--a practical guide to test makers. Princeton, NJ: Educational Testing Service, January 1971.

Branks, J. Proficiency of basic sonar maintenance trainees in the use of common test equipment (NPRA Research Report 66-14). San Diego: U.S. Naval Personnel Research Activity, January 1966.

Brock, J. F., Wells, R. G., & Abrams, M. L. Development and validation of an experimental radiograph reading training program (NPRDC Technical Report 74-33). San Diego: Navy Personnel Research and Development Center, June 1974.

Brown, G. L., Zaynor, W. C., Bernstein, A. H., & Shoemaker, H. A. Development and evaluation of an improved field radio repair course (HumRRO Technical Report 58). Washington, D.C.: Human Resources Research Organization, 1959.

Cogan, E. A., & Lyons, J. D. Framework for measurement and quality control (HumRRO Professional Paper 16-72). Washington, D.C.: Human Resources Research Organization, March 1972.

Cronbach, L. J. Validation of educational measurement. Proceedings of the 1969 Invitation Conference on Testing Problems. Princeton, NJ: Educational Testing Service, November 1969, 35-52.

Crowder, N., Morrison, E. J., & Demaree, R. G. Proficiency of Q-24 radar mechanics: VI. Analysis of intercorrelations of measures (AFPTRC-TR-54-127). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center, 1954.

Dubois, P. H., Teel, K. S., & Petersen, R. L. On the validity of proficiency tests. Educational and Psychological Measurement, 1954, 14, 605-616.

Engel, J. D. Development of a work sample criterion for general vehicle mechanic (HumRRO Technical Report 70-11). Arlington, VA: Human Resources Research Organization, July 1970.

Evans, R. N., & Smith, L. J. A study of performance measures of trouble-shooting ability on electronic equipment. Urbana, IL: University of Illinois, October 1953.

Foley, J. P., Jr. Evaluating maintenance performance: An analysis (AFHRL-TR-74-57(1)). Wright-Patterson Air Force Base, OH: Air Force Human Resources Laboratory, October 1974.

Foley, J. P., Jr. Criterion referenced measures of technical proficiency in maintenance activities (AFHRL-TR-75-61). Wright-Patterson Air Force Base, OH: Air Force Human Resources Laboratory, October 1975.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. Gagne (Ed.), Psychological Principals in Systems Development. New York: Holt, Rinehart, & Winston, 1962.

Highland, R. W. A guide for use in performance testing in Air Force technical schools (ASPRL-TM-55-1). Lowry Air Force Base, CO: Armament Systems Personnel Research Laboratory, January 1955.

Interservice procedures for instructional systems development (NAVEDTRA 106A). Pensacola: Naval Education and Training Command, August 1975.

Jones, A., & Whittaker, P. A validation technique for performance tests. Occupational Psychology, 1973, 47, 189-192.

Klipple, A. G., & Abrams, A. J. Accuracy and consistency in judging active sonar classification cues: IV. Graphic cues (NPRA Technical Bulletin 66-22). San Diego: U.S. Naval Personnel Research Activity, January 1966.

Lennon, R. T. Assumptions underlying the use of content validity. Educational and Psychological Measurement, 1956, 16, 294-304.

Maier, M. H. Performance-based testing in the U.S. Army. Paper presented at 84th Annual Convention of the American Psychological Association, September 1976.

McCluskey, M. R., Trepagnier, J. C., Jr., Cleary, F. K., & Tripp, J. M. Development of performance objectives and evaluation of prototype performance tests for eight combat arms MOSs. Vol. I: Development of performance objectives for eight combat arms MOSs (Final Report-CD-(C)-75-9 Vol. 1). Alexandria, VA: Human Resources Research Organization for U.S. Army Research Institute for the Behavioral and Social Sciences, October 1975. (a)

McCluskey, M. R., Trepagnier, J. C., Jr., Cleary, F. K., & Tripp, J. M. Evaluation of prototype job performance tests for the U.S. Army infantryman (Final Report-CD-(C)-75-9). Alexandria, VA: Human Resources Organization for U.S. Army Research Institute for the Behavioral and Social Sciences, October 1975. (b)

Megling, R. C., & Abrams, M. L. Relative role of experience/learning and visual factors on radiographic inspector performance (NPTRL Research Report 73-22). San Diego: Naval Personnel and Training Research Laboratory, June 1973.

Michaels, W. J., & Karnes, M. R. Measuring education achievement. New York: McGraw-Hill, 1950.

National Occupational Competency Testing Institute-Bulletin of Information for Candidates. Princeton, NJ: Educational Testing Service, 1974.

Osborn, W. C. Process versus product measures in performance testing (Professional Paper 16-74). Alexandria, VA: Human Resources Research Organization, October 1974.

Osborn, W. C. Problems and potentials of applied performance testing. In J. R. Sanders & T. P. Sachse (Eds.). Proceedings of the National Conference on the Future of Applied Performance Testing. Portland, OR: Northwest Regional Educational Laboratory, 1975.

Panitz, A., & Olivo, C. T. National occupational competency testing project. A consortium for occupational competency testing of trade and industrial technical teachers. Phase I: Planning--organizing--pilot testing. Handbook for developing and administering occupational competency tests (Vol. 3). Washington, D.C.: Office of Education (DHEW), February 1971. (a)

Panitz, A., & Olivo, C. T. National occupational competency testing project. Phase II: Directions for area test center coordination, test development, test administration. Washington, D.C.: Office of Education (DHEW), April 1971. (b)

Panitz, A., & Olivo, C. T. National occupational competency testing project. Phase I: Planning--organizing--pilot testing. The state of the art of occupational competency testing (Vol. 2). Washington, D.C.: Office of Education (DHEW), June 1970.

Peach, P. Quality control for management. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1964.

Pickering, E. J. Development of a doppler test (Technical Bulletin 59-27). Washington, D.C.: Bureau of Naval Personnel, October 1959.

Popham, W. J. Performance test of teaching proficiency: Rationale, development, and validation. American Educational Research Journal, January 1971, 8, 105-117.

Ross, R. J. Development of examinations for assessment of occupational competency. New Britain, Conn.: Central Connecticut State College for Connecticut State Department of Education, June 1973.

Saupe, J. L. An analysis of troubleshooting behavior of radio mechanic trainess (AFPTRC-TN-55-47). Lackland Air Force Base, TX: Air Force Personnel and Training Research Center, November 1955.

Schmidt, F. L., Greenthal, A. L., Berner, J. G., Hunter, J. E., Williams, F. M. A performance measurement feasibility study: Implications for manpower policy. East Lansing, MI: Michigan State University for Office of Research and Development, Manpower Administration, U.S. Department of Labor, September 1974.

Shimberg, B., Esser, B. F., & Kruger, D. H. Occupational licensing: Practices and policies. Washington, D.C.: Public Affairs Press, 1972.

Shriver, E. L., & Foley, J. P., Jr. Evaluating maintenance performance: The development and tryout of criterion referenced job task performance tests for electronic maintenance (AFHRL-TR-74-57(II), Pt. 1). Wright-Patterson Air Force Base, OH: Air Force Human Resources Laboratory, September 1974.

Shriver, E. L., & Foley, J. P., Jr. Evaluating maintenance performance: The development of graphic symbolic substitutes for criterion referenced job task performance test for electronic maintenance (AFHRL-TR-74-57(III). Wright-Patterson Air Force Base, OH: Air Force Human Resources Laboratory, November 1974.

Siegel, A. I. Retest-reliability by a movie technique of test administrators' judgments of performance in process. Journal of Applied Psychology, 1954, 38, 390-392.

Steinemann, J. H. Comparison of performance on analogous simulated and actual troubleshooting tasks (PRA Research Memo SRM 67-1). San Diego: Naval Personnel Research Activity, July 1966.

Stuit, D. B. (Ed.). Personnel Research and Test Development in the U.S. Bureau of Naval Personnel. Princeton, NJ: Princeton University Press, 1947.

Swezey, R. W., & Pearlstein, R. B. Developing criterion-referenced tests. Reston, VA: Applied Science Associates, Inc., for Army Research Institute for the Behavioral Sciences, December 1974.

Technical recommendations for psychological tests and diagnostic techniques. Supplement to the Psychological Bulletin. Prepared by a joint committee of the American Psychological Association, American Educational Research Association, and National Council on Measurements used in Education, March 1974.

Vineberg, R., & Taylor, E. N. The interchangeability of job sample tests and job knowledge tests in four army jobs. Paper presented at the 18th Annual Convention of the American Psychological Association, Honolulu, Hawaii, September 1972. (a)

Vineberg, R., & Taylor, E. N. Performance in four army jobs by men at different aptitude (AFQT) levels: Relationships between performance criteria (HumRRO Technical Report 72-23). Arlington, VA: Human Resources Research Organization, August 1972. (b)

Vineberg, R., & Taylor, E. N. Manual for developing skill qualification tests. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, in press.

Whipple, J. E., Baldwin, R. D., Mager, R., & Vineberg, R. A three-hour performance test to evaluate job effectiveness of army radar mechanics (HumRRO Professional Paper 12-69). Washington, D.C.: Human Resources Research Organization, 1969.

Williams, W. L., Jr., & Whitmore, P. G., Jr. The development and use of a performance test as a basis for comparing technicians with and without field experience: The NIKE AJAX AFC maintenance technician (Technical Report 52). Washington, D.C.: Human Resources Research Office, January 1959.

Wilson, C. L. On-the-job and operational criteria. In R. Glaser (Ed.). Training Research and Education. Pittsburgh, PA: University of Pittsburgh Press, 1962.

Wilson, C., & Mackie, R. Research on the development of shipboard performance measures: Part I--The use of practical performance tests in the measurement of shipboard performance of enlisted personnel. Los Angeles: Management and Marketing Research Corporation for the Office of Naval Research, November 1952.

Wilson, C. L., Mackie, R. R., Buckner, D. N., Seigel, A. I., & Courtney, D. A manual for use in the preparation and administration of practical per formance tests (NAVPERS 91961). Washington, D.C.: Bureau of Naval Personnel, undated.

Winchell, J. D., Panell, R. C., & Pickering, E. J. A personnel readiness training program: Operation of the AN/BQR-20A (Tech. Rep. 77-4). San Diego: Navy Personnel Research and Development Center, December 1976.